Differentially private de-identifying textual medical document is compliant with challenging NLP analyses: example of privacy-preserving ICD-10 code association

Yakini Tchouka[a], Jean-François Couchot[a], David Laiymani[a], Philippe Selles[b] and Azzedine Rahmani[b]

[a] Femto-ST Institute, Univ. Bourg. Franche-Comté, CNRS, France
[b] Hôpital Nord Franche-Comté (HNFC)

ARTICLE INFO

Keywords:
De-identification
Clinical data
Local Differential privacy
metric-privacy
Natural language processing
ICD-10 code association
Machine learning

Abstract

Medical research plays a crucial role within scientific research. Technological advancements, especially those related to the rise of machine learning, pave the way for exploring medical issues that were once beyond reach. Unstructured textual data, such as correspondence between doctors, operative reports, etc., often serves as a starting point for many medical applications.

However, for obvious privacy reasons, researchers do not legally have the right to access these documents as long as they contain sensitive data, as defined by regulations like GDPR or HIPAA. De-identification, meaning the detection, removal or substitution of all sensitive information, is therefore a necessary step to facilitate the sharing of these data between the medical field and research. Over the last decade, various approaches have been proposed to de-identify medical textual data. However, while entity detection is a well-known task in the natural language processing field, it presents some specific challenges in the medical context. Moreover, existing substitution methods proposed in the literature often pay little attention to the medical relevance of de-identified data or are not very resilient to attacks.

In this paper, we delve into these challenges. Firstly, we implemented an efficient system for detecting sensitive entities in French medical data to subsequently substitute them accurately. Secondly, we provided robust strategies for generating substitutes that incorporate the medical utility of the data, thereby minimizing the utility difference between the original and de-identified data, and that mathematically ensure privacy protection. Thirdly, we evaluated the utility of the de-identification system in a context of ICD-10 code association. Finally, we presented various systems developed to tackle ICD-10 code association while providing a state-of-the-art model in French.

1. Introduction

Artificial Intelligence (AI) is prevalent across diverse domains encompassing finance, transportation, information management, and beyond (to mention only a fraction). The sphere of healthcare is by no means an exception to this pervasive trend. This influence extends even to the manipulation of unstructured data, such as textual information. Prediction that were insurmountable are now emerging as tractable challenges. Examples of these encompass tasks like the identification of similar patient records, automatic ICD labelization [AND+19, NRG+18, MDS23, PTM+23, TCL+23b], the anticipation of hospital re-admissions [ABBZ17, TST22], healthcare-associated infections detection [TKB+18, SBS+20], to mention just a few.

Nonetheless, primarily as a consequence of resource and time constraints within healthcare institutions, the execution of these processes necessitates the involvement of computer experts specializing in deep learning, data science, and big data analytic. Consequently, it becomes imperative to facilitate the exchange of data between healthcare practitioners and data science specialists. Given the sensitive and critical nature of the data in question, such collaboration mandates a rigorous de-identification procedure, which can only be conducted within a legal framework that governs the stakeholders within the healthcare domain. Notable examples of institutional regulations include the U.S. Health Insurance Portability and Accountability Act (HIPAA) [CM18] and the European General Data Protection Regulation (GDPR) [EU16].

These regulations specifies that anonymising data is a prerequisite for sharing it. When processing tabular data, each record contains values for a set of collected attributes. Numerous methods have addressed this or that type of attribute [EPK14, DKY17] and their combinations in an array [Swe02, WXY+19]. However, when dealing with unstructured data, the de-indentifiability problem is more complex, since we have to work with an a priori unorganised set of information. But it is in fact organised (the words form sentences) and this organisation must be preserved. Indeed, medical information after de-identification must be as rich as it was before de-identification in order to hope for subsequent processing by AI, as in the examples mentioned above.

This study is centered on the capability to exchange textual medical documents, often authored by healthcare professionals, which can manifest as surgical reports, clinical notes, or results from biological examinations.

To enhance privacy protection, de-identification techniques [DLUS16, HHD+20, BAC+21, HSDF21, LPS+21, JFHS22] have been put forth as a means to mask or substitute any form of Protected Health Information (PHI) attributed to a patient, rendering it arduous to associate an individual with their spe-

ORCID(s):

cific data. The components constituting PHI are partly delineated by the privacy regulations in force within the relevant jurisdiction. As an illustrative example, HIPAA provided 18 distinct categories of PHI, encompassing designations such as names, geographic locations, and telephone numbers. In the European context, given that GDPR does not explicitly offer PHI definitions, the majority of researches adhere to the stipulated HIPAA criteria.

The de-identification process is typically depicted as an algorithm comprising two primary steps, namely first, identification of all sensitive information, and, next, replacement of these identified elements.

The initial step encompasses the automated detection of entities within a given text sequence. In the realm of Natural Language Processing (NLP), this phase is commonly referred to as Named Entity Recognition (NER). This task presents however several unique challenges in a medical context among of them is the scarcity of a consistent training dataset in the specific language relevant to the effective model development (which is essentially the dataset that the de-identification task seeks to generate). Another notable challenge lies in the need to fine-tune and adapt generic natural language processing models to the specialized lexicon and terminology prevalent in the field of medicine.

The subsequent step involves the substitution of the identified sensitive entities. Regrettably, this step is frequently overlooked in research efforts, as it is perceived as less critical. This omission is particularly evident when dealing with entities like phone numbers or email addresses, which can be replaced with arbitrary random values, given their lack of direct correlation with medical information. However, the same does not hold true for entities such as dates or geographic locations. For instance, analyzing the medical records of a patient under 18 differs significantly from analyzing those of a patient over 50. Similarly, the patient's location, whether in an area with high pollution levels or a region with low pollution, has notable implications for medical assessment. The challenge in this second step lies in striking a delicate balance between safeguarding individual privacy and preserving the medical utility of the data, which remains essential for subsequent machine learning tasks.

This study is motivated by the objective of developing a Machine Learning approach aimed at assigning numerical codes from the International Classification of Diseases, 10th Revision (ICD-10), to medical documents such as patient records, medical reports, and clinical notes. The ICD-10 is a globally recognized standard classification system that assigns numeric codes to various medical conditions, symptoms, medical procedures, and other health-related information. The task of associating ICD-10 codes with medical documents presents several significant challenges, including the extensive number of codes to be classified, the scale of the input data, language variations (e.g., French versus English) in the automatic natural language processing models, imbalanced datasets, and more.

However, to develop this Machine Learning approach, access to data is required, and this is where the entire de-identification process becomes significant: enabling prototyping on data sanitized of identifying information while preserving a substantial portion of medical information.

This article constitutes a comprehensive reworking of articles [TCC$^+$22, TCL23a, TCL$^+$23b], demonstrating the feasibility of establishing a comprehensive process for de-identifying medical notes based on Local Differential Privacy.

Our contributions can be summarized as follows:

1. We developed a sensitive French Named Entity Recognition system based on the Transformer architecture (FlauBERT model). This model is trained on a constructed dataset and represents the state-of-the-art in named entity recognition in the French language within the context of de-identification with HIPAA attributes. An open-source implementation of this NER task is available on GitHub [1]
2. We formalized robust strategies for generating substitutes for sensitive attributes based on Local Differential Privacy based on a metric (metric-privacy) that ensures security and preserves medical utility. An open-source implementation of the surrogate generation approaches provided in is available on GitHub[2].
3. These initial two contributions enabled the development of an incremental approach for dataset construction: from a partially manually labeled dataset of 375 individual records to a de-identified dataset containing over 46,000 documents, in accordance with the criteria outlined earlier.
4. We developed various architectures to tackle issues of ICD coding, namely the significant number of labels and the size of medical data. The proposed "CamemBERT+LAAT" architecture represents the state-of-the-art in the French language for ICD-10 code association task. An open-source implementation of this system is available on GitHub[3].
5. We evaluated our de-identification approach in the context of ICD-10 code association task. Our de-identification approach helps reduce the loss of utility (6.8%) compared to a traditional de-identification method (12%, e.g., replacing sensitive attributes with their labels).

This work is conducted in collaboration with the Hopital Nord Franche-Comté (HNFC), a French public health center that generously provided us with patient stay records. To uphold privacy standards, all experiments were conducted on-site, and no data was extracted from the hospital premises.

This paper is organized as follows. Section 2 presents the state of the art in the various domains we have addressed, namely, Natural Language Processing, de-identification, and the association of ICD-10 codes. Subsequently, we discuss in Section 3 the de-identification with the Named Entity Recognition task (Section 3.1), followed by substitution (Section 3.2). Next, in Section 4, we delve into the application context

[1] https://github.com/mlfiab/ner-french

[2] https://github.com/mlfiab/surrogate-deid

[3] https://github.com/mlfiab/icd10-french

		[TCC+22]	[TCL23a]	[TCL+23b]	PROPOSAL
De-identification	Named Entity Recognition	Hybrid Method • MEDINA • Flaubert-ner	Deep Learning only • new labelled dataset for NER • Flaubert-ner only	NA	• Rewriting • Publication on GitHub
	Surrogate generation	ϵ-LDP based • Date: bounded Laplace mechanism • Location: GeoIndisdinguishability	ϵ-metric privacy based • Date: metric Laplace mechanism • Location: exponential mechanism	NA	• Rewriting, Formalizing • Publication on GitHub
ICD-10 association	Datasets	128 documents • 64 original ones • 64 de-identified ones	NA	ICD-10-HNFC • 56, 000 documents • with ICD-10 associated codes	Deidentified dataset • ICD-10-DEID-HNFC, ICD-10-TAG-HNFC • with ICD-10 associated codes
	Evaluation method	Manual task • ICD-10 association on all documents • manual confusion matrix computation	NA	• CamemBERT + LAAT • Bibliographic comparison with state of the art	• Model improved thanks to de-identified datasets • Rewriting of [DCC+20] model • Fair experimental comparison

Table 1
Comparison of main contributions of articles [TCC+22, TCL23a, TCL+23b]

of this paper, the association of ICD-10 codes. Then, Section 5 presented the methodology to measure the utility of de-identification in the context of ICD-10 code association. Section 6 presents all the experiments conducted within the scope of this work. Finally, Section 7 discusses the pertinence of our approaches in the time of LLMs and Section 8 presents our concluding remarks.

2. Related Work

In this section, we present the state of the art in the various domains we have addressed, namely, Natural Language Processing, de-identification, and the association of ICD-10 codes.

2.1. Natural Language Processing

NLP has undergone significant advancements in recent years, primarily driven by the introduction of the Transformers model [VSP+17]. These models exhibit a remarkable capacity for transfer learning, as demonstrated by the BERT model [DCLT18], which have proven their effectiveness in providing more accurate contextualized representations. Subsequently, a range of pre-trained models emerged, including BERT, RoBERTa [LOG+19], and others. These models undergo pre-training on extensive general domain of a given language text to capture text data modeling capabilities. They are then fine-tuned for specific classification tasks.

In the realm of French language processing, two prominent models have been introduced: FlauBERT [LVF+19] and CamemBERT [MMOS+20]. Additionally, there are multilingual models like XLM-R [CKG+19]. Some models specialize in domain-specific text, ClinicalBERT [AMB+19] and BioBERT [LYK+20] for instance, which are trained on medical data for medical domain tasks. Notably, there's a gap in French language models, hindering the application of machine learning approaches to French documents compared to English ones. In the French language, we can mention DrBERT, introduced recently by [LBD+23], trained on a large corpus of medical data in the French language and based on the BERT architecture.

Transformers models generally have a limited input size, typically around 512 tokens in practice. However, this limitation becomes especially problematic for clinical documents, which are often much larger than 512 words or tokens. In addressing this challenge, a hierarchical approach was proposed in [PZV+19]. The document is segmented, and each segment is processed by a Transformers model. The encoded segments are then aggregated in the next layer, employing techniques such as linear layers, recurrent neural networks, or other layers of Transformers. In recent years, a sparse-attention system known as the *LongFormer* model has been introduced in [BPC20]. This model incorporates both local attention (within a window of neighboring tokens) and global attention, effectively reducing the computational complexity. Consequently, the LongFormer model can handle document sizes of up to 4096 tokens.

Recently, NLP has experienced a major breakthrough with the introduction of new learning approaches using Large Language Models (LLM), such as Reinforcement Learning from Human Feedback (RLHF) [MHL+17] on GPT [RWC+19]. With these LLM, training techniques like zero-shot learning [AHL+22] and few-shot learning [SKW+22] make it easier to approach NLP tasks.

2.2. De-Identification

2.2.1. *Named Entity Recognition*

Detecting entities in a text sequence is a well-established task in automatic natural language processing, commonly referred to as Named Entity Recognition (NER). Initially, rule-based systems were employed, where a set of rules covered all the potential syntaxes of the attributes to be identified. With the advent of machine learning, more precise systems emerged, utilizing models such as SVM, Decision trees, or Conditional Random Field (CRF). The transition to neural networks saw the introduction of deep learning models by researchers like Dernoncourt [DLUS16] and Liu [LTWC17]. Dernoncourt's recurrent neural networks achieved impressive F_1-scores of 99.23% and 97.85% on i2b2 [SRU13] and MIMIC [JPS+16a] datasets, respectively, while considering attributes outlined by the HIPAA law [CM18].

With the rise of transformers [VSP+17] and BERT in NLP, which excel in contextualized text representation, it has become evident that transformer-based models are the most accurate for NER. Noteworthy works in this domain include [Han21] and [PdGS21]. In a recent study [LYZ+23], GPT-4 with the "zero-shot learning" technique [PAL+22] was employed for NER in the context of de-identification, achieving a remarkable F_1 score of 99% on the i2b2 dataset. In comparison, models using BERT and ClinicalBERT achieved 79.8% and 97.4%, respectively.

Research on the de-identification of French medical documents, led mainly by [GGN15], utilized a CRF-based machine learning model that reached an 80% F_1-score. A recent work [BAC+21] focuses on de-identifying French emergency medical records. It employs a two-step approach: the FlauBERT model assigns labels to documents requiring de-identification, followed by a combination of rules-based techniques and LSTM through Flair [ABB+19].

Unfortunately, there is no French equivalent to datasets like MIMIC or i2b2. This led authors in [TCC+22] to combine MEDINA a CRF based machine learning method [GGN15] and a neural network method based on transformers on the WikiNER dataset [NRR+13] to integrate all attributes for detection. This hybrid system achieves a baseline F_1-score of 94.7% in this context.

Note that the use of multilingual models such as XLM-Roberta [CKG+19] can be an interesting approach when the datasets are scarce. In [SGCP20] the authors show that such models can be useful. Nevertheless, in [KCG+23], we have tested this approach on a similar problem of classifying emergency phone calls and the results we obtained were not as good as the full French approach i.e. CamemBERT or FlauBERT with a small French dataset .

2.2.2. *Entity Substitution*

Once the entities have been detected, they need to be cleaned up - this is the substitution phase. The simplest consists in replacing the detected information by its entity (the name "Durant" by the entity "NAME"). While this protects the privacy of individuals, it degrades the structure, readability and coherence of the document. For this reason, more coherent de-identification methods have emerged [SUK+15, DLN+14, ULS07]. Sensitive information is replaced by random information that preserves the document's structure (the name "Durant" is replaced by a random name "Julien", or a date by a date shifted by a few days, etc.). In these strategies, the aim is to preserve the document's structure and information while protecting privacy. In the medical context, the task of substitution can become complex to tackle. In general, in medical analyses, some of the sensitive attributes to be anonymized may have an impact on document analysis, such as ages, dates or geographical locations.

In [TCC+22], for temporal entities, the authors opted for Local Differential Privacy [DJW13] with bounds in time categories to calibrate the added noise. Regarding geographical locations, geo-indistinguishability [BCP14] was retained as a direct mechanism to provide a location close to the original one. While these two privacy-preserving methods effectively safeguard the data privacy, it is important to note that they may introduce unnecessary additional noise in the temporal aspects, and in some cases, they could potentially substitute a large urban area with a small village. This substitution may result in a significant disparity in the epidemiological characteristics of the data, as the characteristics of a small village could be vastly different from those of a city.

To address these issues, [TCL23a] have proposed new strategies. For temporal data, the authors used differential privacy based on a metric with Laplace's mechanism. This allows the distance between dates to be taken into account in the substitution. For geographical locations, the authors integrated health criteria to establish a vector distance between locations. In the context of differential privacy based on this distance, the exponential mechanism is used to substitute elements.

In [HSDF21], the authors use ontology based generalization techniques such as YAGO [FGG+07] to substitute detected entities and security is observed through evaluating robustness against re-identification attacks. Finally, in articles [FBDD20, CVFW23], the authors present an approach adding noise verifying ϵ metric-privacy to vector representation of words resulting from word embedding models. However, this approach in one step is not compliant with HIPAA constraints.

2.3. ICD Code Association

The automatic association of ICD codes stands as a prominent challenge in medical research, and with the advancements in neural networks and the evolution of natural language processing, numerous approaches have been explored. Authors such as [CBS+16] and [BNKC+18] utilized recurrent neural networks (RNNs) to encode Electronic Health Records (EHR) for predicting diagnostic outcomes. Conversely, [SXH+17] and [MWD+18] incorporated attention mechanisms with RNNs and CNNs to enhance model accuracy.

[XX18] and [TCC19] introduced diverse strategies for considering the hierarchical structure of codes. [XX18] employed a sequence tree LSTM to capture the hierarchical relationships and semantics of each code, while [CCL+20] proposed training the integration of ICD codes in a hyperbolic space, leveraging a graph neural network to capture code co-occurrences. LAAT [VNN20] integrates a bidirectional LSTM with an attention mechanism incorporating labels.

EffectiveCAN [LCK+21] introduced a squeeze-and-excitation network, residual connections, and representation extraction from all encoder layers for label attention. They addressed long-tail prediction issues with focal loss, achieving an F_1-score of 58.9% on MIMIC 3 [JPS+16b]. ISD [ZCC+21] utilized shared representation extraction and a self-distillation learning mechanism to address the distribution of long-tailed codes with 55.9% of F_1-score.

[PTM+23] proposes models based on neural networks using a supervised learning system. The authors developed convolutional neural network (CNN) with PubMedBERT embeddings, using a dataset of 15,329 medical documents. They validated their approaches on the MIMIC 3 dataset with an 55.65% F_1-score. [HTC22] proposed the PLM-ICD system, focusing on document encoding with multi-label classification. They employed a transformer-based encoding model adapted to medical corpora. To handle a large set of labels (e.g., MIMIC 3 with over 8,000 codes), they utilized the Label-Aware Attention (LAAT) mechanism [VNN20] and implemented a hierarchical method for long sequences. PLM-ICD is the current state-of-the-art model, achieving 59.8% F_1-score on MIMIC 3 [JPS+16b] and 50.4% on MIMIC 2 [SVR+11].

In the French context, [DCC+20] proposed Convolutional Neural Networks (CNN) models with multi-label classification system for ICD-10 code association. They utilized FastText vectors [BGJM16] for document encoding, considering both the final labels and grouped families to reduce the class count. Trained on a private dataset of 28,000 clinical documents, their model achieved an F_1-score of 39% with 6,116 codes and 52% with 1,549 codes. In [MDS23], you can find a summary of the state-of-the-art of the ICD coding models we just mentioned. In [TCL+23b], the authors utilized a multi-label classification system as in [HTC22]. They experimented with various architectures based on transformers models to address the challenges posed by a significant number of codes and the size of input sequences. The authors propose a code association system based on the most frequent codes. The most advanced model from [TCL+23b] currently represents the state-of-the-art in the task of associating ICD-10 codes in the French language.

3. De-identification System

In this section, we discuss the de-identification process. We present the two de-identification steps (named entity recognition and entity substitution) and our proposals in each of them. The following is our guiding example that will be developed throughout this section.

Thread Example. Consider the following fictional sentence. It is typical of what can be present in a medical text document of a hospital. "Mr. Durand born in Dijon, 40 years old, was admitted to the hospital from 12/02/2020 to February 26, 2020 following a road accident in Dijon".

3.1. Named Entity Recognition

In section 3.1.1 we discuss the motivation for the need to strengthen the NER step. In section 3.1.2, we present the methodology for building the training dataset. Finally, in section 3.1.3, we provide details on the architecture of our NER model and the training process.

Thread Example. Figure 1 illustrates the result of a perfect detection (NER) process applied to the threaded example. HIPAA labels [CM18] with their descriptions are summarized in Table 2.

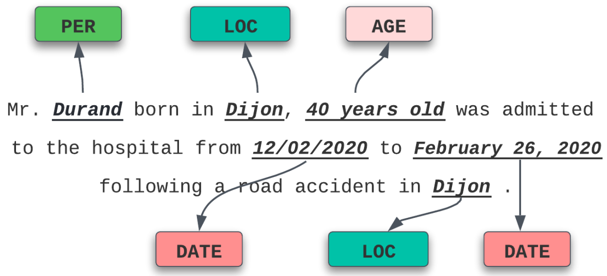

PER
LOC
AGE
Mr. Durand born in Dijon, 40 years old was admitted
to the hospital from 12/02/2020 to February 26, 2020
following a road accident in Dijon .
DATE
LOC
DATE

Figure 1: Perfect NER of PHI entities on thread example.

3.1.1. *Motivation and Global Overview*

Achieving near-perfect scores in Named Entity Recognition (NER) is crucial for the successful de-identification process. Undetected sensitive information poses a risk of document re-identification. NER tasks are problem-dependent,

Label	Description
PER	All names of persons
DATE	All date sequences in all formats
LOC	All geographical locations and zip codes
ORG	Organizational entities
AGE	Ages
TEL	Phone Numbers
REF	All references related to individuals
QID	Any ID sequence

Table 2
Description of sensitive attributes based on the HIPAA law.

often lacking the ideal dataset for specific issues. The most successful NER results in de-identification, as seen in literature, have been achieved by implementing English models on domain-specific datasets that are both consistent and comprehensive. For the French language, datasets with sufficient data to build an efficient model are rare, and as far as we are aware, none exists in the medical domain. Consequently, it becomes imperative to construct a suitable dataset tailored to our application's context—a medical corpus containing all pertinent identifying attributes. With access to such a dataset, we can leverage a Transformer-based NER method.

3.1.2. *Building a Labelled Dataset*

As mentioned, the most challenging step involves finding a sufficiently large dataset for implementing an accurate model that encompasses all categories of sensitive information and is tailored to the medical corpus, especially in French. Unfortunately, such a dataset is currently unavailable or not easily accessible to everyone. In [TCC$^+$22], the authors faced similar challenges, using the WikiNER dataset, which had only a few tags and a very general vocabulary. To address this limitation, they had to employ multiple methods to integrate all categories into a de-identification tool. In this work, conducted in collaboration with a French public hospital, we have on-site access to a substantial set of unlabeled medical notes. Our proposal involves members of the hospital to semi-manually annotate a subset of these notes.

The process is illustrated in Figure 2. Given that there may be a necessity to access these documents in subsequent stages, this entity recognition process begins with deploying the de-identification step outlined in [TCC$^+$22]. Consequently, an automated pre-annotation phase is executed using the same tool presented in [TCC$^+$22]. This step serves merely an efficiency purpose. Subsequently, a third phase involving human verification, correction, and enhancement of these annotations is conducted, which constitutes the most crucial step of this process. Thanks to this pre-annotation phase, the entire annotation process took only 25 hours for one individual (equivalent to 1 minute per file). The ensuing dataset, labeled as *FrenchHospitalNER*, is thus de-identified according to the methodology described in [TCC$^+$22] and includes tags suitable for Named Entity Recognition (NER) based on the Transformer model. It comprises 14,925 sentences.

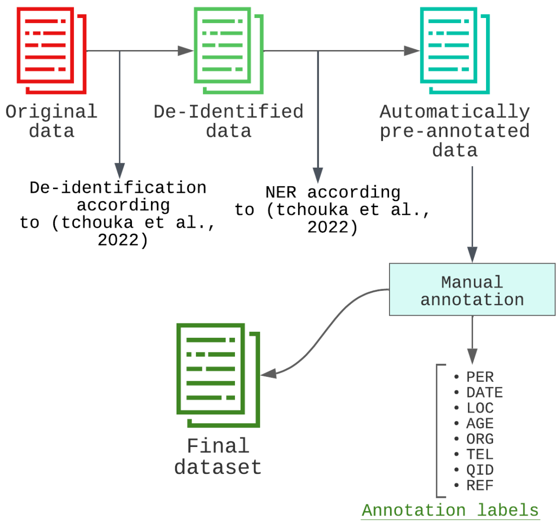

Original data
De-Identified data
Automatically pre-annotated data
De-identification according to (tchouka et al., 2022)
NER according to (tchouka et al., 2022)
Manual annotation
Final dataset
• PER
• DATE
• LOC
• AGE
• ORG
• TEL
• QID
• REF
Annotation labels

Figure 2: FrenchHospitalNER Dataset Construction.

3.1.3. *Supervised Learning on a Dedicated Labelled Dataset*

This section starts with the introduction of our model's architecture. Subsequently, we delve into the implementation details of supervised learning. Finally, we present an evaluation of our model, drawing comparisons with existing de-identification models.

Model Architecture: Transformer Based Approach

The accessibility of FlauBERT, a BERT-based pre-trained French model, has influenced our decision to use this transformer. This choice is further supported by findings in [PdGS21], demonstrating that language-specific models for French, such as FlauBERT, yield improved results compared to multilingual BERT models.

Fine-tuning Transformer Model

Starting with a pre-trained model like FlauBERT, the remaining task involves fine-tuning it on a smaller and more specialized dataset. Instead of creating our text classification or feature detection model from scratch, our approach involves starting with the pre-trained BERT and adding a classification layer i.e. a fully-connected layer. This is the transfer learning process introduced by [HR18]. This process is illustrated in Figure 3.

Training

The learning process has been executed using the aforementioned dataset, employing a deep learning-based NLP model as described earlier.

- the **Learning Rate** controls the size of the update steps along the gradient. Usually, a very small value is set (10^{-4} in this work), so that the weights are less modified at each iteration, which avoids missing the optimal values of the error function

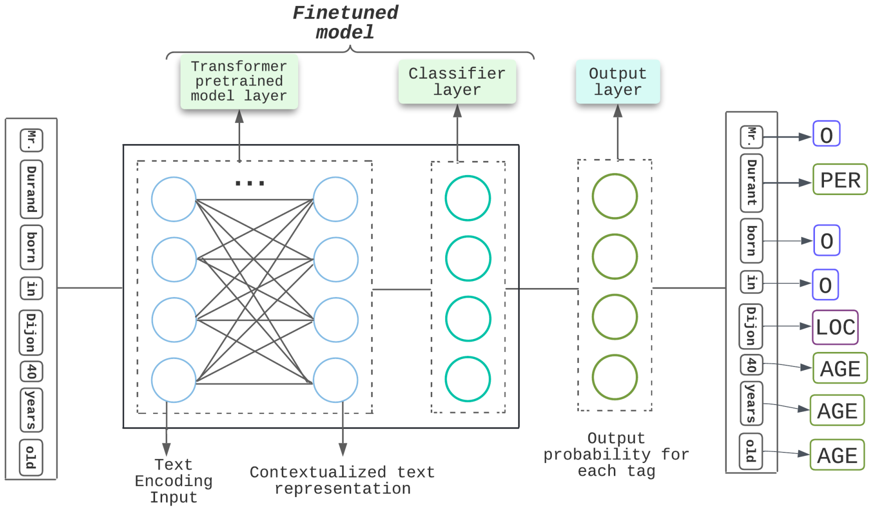
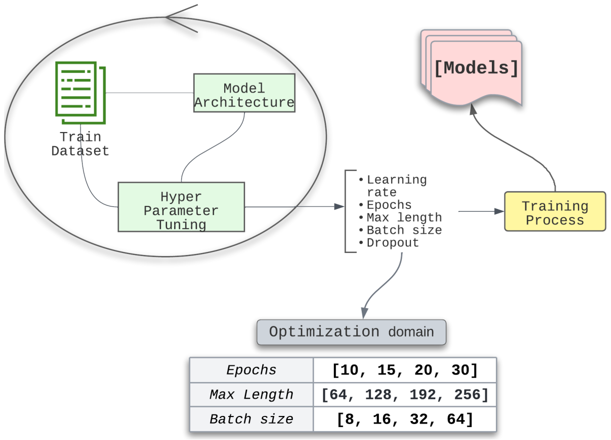


Figure 3: Deep Learning Model Architecture for NER.

- the **Dropout** is a regularization technique for reducing overfitting in neural networks. It is set to 0.1 in this work, which means that 10% of selected neurons are ignored during training
- the **Training Batch Size** is the number of training samples to work through before the model's internal parameters are updated.
- the **Maximum length** defines the maximum number of words in the sentences
- the **Number of epochs** is the number of complete passes through the training dataset.

The Named Entity Recognition (NER) model is formulated as a multi-class classification task, where tags serve as classes. The CrossEntropy error function [DBKMR05] is well-suited for this purpose. For optimization through backpropagation, the AdamW algorithm [LH17], one of the latest advancements in optimizers known for its effectiveness in neural network learning, is employed. The FrenchHospitalNER dataset is randomly divided into training and test sets at a ratio of 90/10.

In the realm of machine learning, the challenge lies in determining the optimal values for hyper-parameters to achieve the most accurate results. Various methods exist for hyper-parameter optimization, including Grid or Random Search, model-based Bayesian methods. . . Studies, such as [BYC13], indicate that Bayesian methods tend to yield more accurate results. In this paper, the Tree-structured Parzen Estimator [BBBK11], a classical Bayesian optimization algorithm suitable for a classification model like ours, was employed for hyper-parameter optimization. We conducted experiments with different parameter combinations based on the Tree-structured Parzen Estimator algorithm, as illustrated in Figure 4.



Epochs	[10, 15, 20, 30]
Max Length	[64, 128, 192, 256]
Batch size	[8, 16, 32, 64]

Figure 4: Hyper-parameter optimization process. Search domains are provided for number of epochs, maximum length, and batch size.

The training optimization process focused on three key hyper-parameters: the number of epochs, the maximum text length, and the batch size. Among all the hyper-parameter configurations, the one with highest F_1 score is *number of epochs* = 20, *maximum length* = 128, and *batch size* = 64. The model with such a configuration is further utilized in the next Evaluation section.

3.1.4. Evaluation

To evaluate our model, we used the classical metrics: Precision (P), Recall (R), and F_1-score[4] which is an harmonic mean of both.

To get a sense of the overall performance of the system, we use the micro-average system. To be fair with [TCC+22], HNFC hospital provides the same validation set already presented in [TCC+22] that will refer to as HNFC dataset.

[4] https://en.wikipedia.org/wiki/F-score

HNFC dataset contains 375 documents of deceased patients from HNFC hospital. These documents were pre-annotated by the hybrid system proposed by [TCC+22] and then were manually annotated by the hospital staff with Doccanno manual annotation tool [NKK+18]. Given the automatically pre-annotated files, the annotator is responsible for checking, correcting and completing the possible errors produced by the model. HIPAA attributes were formalized to avoid ambiguity and multiple annotation criteria. To minimize the risks, two annotators worked in parallel on the same files and each annotation pair is then manually analyzed and merged into an unique one. This work was performed by 6 people during 6 hours. This process is illustrated in 5.

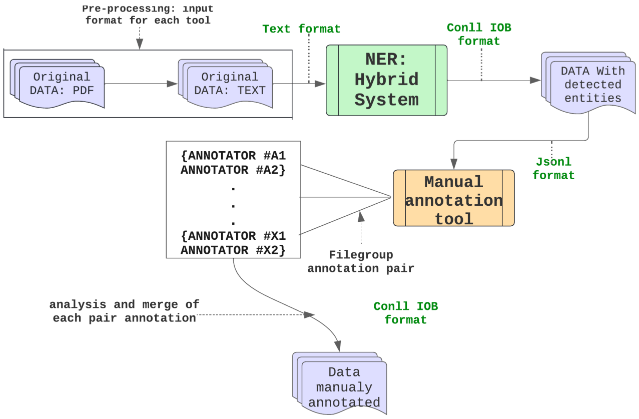

Pre-processing: input format for each tool
Original DATA: PDF
Original DATA: TEXT
Text format
NER: Hybrid System
Conll IOB format
DATA With detected entities
Jsonl format
Manual annotation tool
{ANNOTATOR #A1 ANNOTATOR #A2}
.
.
.
{ANNOTATOR #X1 ANNOTATOR #X2}
Filegroup annotation pair
analysis and merge of each pair annotation
Conll IOB format
Data manualy annotated

Figure 5: HNFC Dataset Construction.

The i2b2 dataset [SRU13] contains over 20,000 medical labeled reports allowing examination of diverse biomedical topics like automatic assignment of ICD-10 codes (as discussed later). When used by Dernoncourt et al [DLUS16], only 40% of it is taken into consideration.

To establish a baseline, we evaluated various existing entity detection solutions on the same evaluation dataset. Here are the baseline models we integrated into our study:

- CamemBERT-ner [Pol] is a NER model that was fine-tuned from camemBERT [MMOS+20] on WikiNER dataset [NRR+13],
- MEDINA [GGN15] is a method that employs CRF to label entities such as dates, phone numbers, email addresses, and URLs in medical texts written in French language,
- FlauBERT-ner: A model built on deep learning from a pretrained transformers model, FlauBERT, also trained on Wiki-NER.

The results are detailed in Table 3. In this table, the highest scores in any language (resp. in French language only) for each entity (PER, ORG, LOC…) with respect to Precision, Recall and F_1 metric are marked in **bold**, (resp. in *italic*). In French language, this proposal significantly outperforms all the approaches for almost all entities and for the micro-average aggregation. The enhanced performance can be attributed to the incorporation of a BERT-based layer, enabling precise contextualization of the sequence. Compared to English evaluation (on two distinct dataset), the results of this proposal are not as far from those presented in [DLUS16], which is recognized as the state of the art. Our comparatively lower score in the organization category, when contrasted with the i2b2 model, can be attributed to the informal structuring of organizations in medical documents, often in the form of abbreviations or isolated words. Increasing the dataset is expected to mitigate such challenges and contribute to overall improvements in scores across different categories.

The next step is substituting the detected entities, as described in the next section.

3.2. Surrogate Generation Strategies

The challenge here lies in replacing personally identifiable information (PII) detected by Named Entity Recognition (NER) with relevant surrogates related to medical content while preserving privacy.

In the objective of preserving medical information, all the entities do not have the same need of attention. For instance, replacing a phone number by any other random phone number does not change the medical information. On the opposite side, replacing location and date entities with random ones may lead to loose geographic epidemiological data for the former and chronological aspect for the latter.

For name elements, a random and memoïzation [EFM+19, ACABX22] based algorithm preserving the affiliation within the documents is provided in [TCC+22] and is kept here.

Thread Example. In the thread example, "Durand" could be replaced by any name, for example, "Julien."

In contrast, temporal and location data inherently carry information that is both medically critical and highly identifiable. This work will focus on these attributes, sanitized with Local Differential Privacy based approach first recalled hereafter. Local Differential Privacy [DJW13] formalizes the algorithm's robustness, and its definition is recalled in next section.

3.2.1. *Local Differential Privacy (LDP)*

An algorithm adhering to the principles of Local Differential Privacy (LDP) should be applied before granting access to the data curator, either by the data owner themselves or as a pre-treatment input for the data curator.

Such an algorithm ensures that when a data owner publishes a sanitized output, it does not modify at all (from a probabilistic point of view) the ability to distinguish between him/her and the other data owners. As a consequence, the wider the range of possible values (called sensitivity), the greater the amplitude of the added noise.

Definition 1 (ϵ-local differential privacy)**.** *A random mechanism $\mathcal{M}$ satisfies ϵ-local differential privacy if, for any pair*

Method	CamemBERT-ner			MEDINA			FlauBERT-ner			**[TCC+22]**			**PROPOSAL**			[DLUS16]		
Dataset	*HNFC*															i2b2		
Metric	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
PER	89	99	93.8	***98.2***	97.7	*98.2*	91.8	97.6	94.6	96.3	***99.8***	98	97.2	98.9	98	**98.2**	99.1	**98.6**
ORG	7.	21.8	11.1	32.6	24.8	28.1	16.9	34.1	22.6	41.1	*57.3*	47.8	*90*	51	*65.6*	**92.9**	**71.4**	**80.7**
LOC	46	67.2	54.6	98.8	81.1	89.1	75.7	66.3	70.7	88.4	***95.8***	92	***99.4***	94.4	***96.9***	95.9	95.7	95.8
DATE	NA			97.7	86.6	91.9	NA			97.7	86.7	91.9	***99.2***	*95.7*	*97.4*	99	**99.5**	**99.2**
AGE	NA			91.5	66.9	77.3	NA			91.5	66.9	77.3	*98.2*	*91.8*	*95*	**98.9**	**97.6**	**98.2**
TEL	NA			99.5	97.9	98.7	NA			***99.5***	97.9	98.7	99.4	***99.8***	***99.6***	98.7	99.7	99.2
REF	NA			NA			NA			NA			96.1	79.5	87	NA		
QID	NA			NA			NA			NA			77.2	32	45.3	**99.2**	**98.7**	**99**
Mic.-avg.	70.8	51.5	59.6	98.2	91.2	94.5	85.8	86.7	86.3	94.6	94.9	94.7	***98.5***	*96.4*	*97.4*	98.3	**98.5**	**98.4**

Table 3
Comparison of NER metric results across diverse entities. The highest scores in any language (resp in French language only) for each entity with respect to Precision, Recall and F_1 metric are marked in **bold**, (resp. in *italic*). NA stands for not addressed entity.

of input values v_1, v_2 in the domain of $\mathcal{M}$ and any possible output y of $\mathcal{M}$:

$$\Pr[\mathcal{M}(v_1) = y] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(v_2) = y]. \tag{1}$$

LDP mechanisms [YLZ+20] are calibrated based on the data types they handle, which include real and integer values, among others. They are tailored to address various questions, such as frequency estimation, the identification of heavy hitters, and joint distribution analysis. One prominent mechanism among these is the Laplace mechanism, which is directly applicable to numerical data.

Definition 2 (Laplace mechanism in an interval of amplitude Δ)**.** *In the Laplacian mechanism, a numerical value v in an interval of of amplitude Δ is sanitized into a numerical value*

$$\mathcal{M}_{\text{Lap}}(v, \Delta, \epsilon) = v + \text{Lap}\left(\frac{\Delta}{\epsilon}\right) \tag{2}$$

where $\text{Lap}\left(\frac{\Delta}{\epsilon}\right)$ is the Laplace distribution centered in 0 and whose scale parameter is $\frac{\Delta}{\epsilon}$.

Laplace mechanism is however not adapted if it is required to return a precise answer (*i.e.* without noise, location name for instance). One solution is the exponential mechanism[MT07] recalled here which is going to select the sanitized output from a provided set element whilst preserving ϵ-local differential privacy.

Definition 3 (Exponential mechanism [MT07])**.** *Let v in a domain $\mathcal{D}$ be the value to sanitized, $\mathcal{R}$ be the set of possible output sanitized data and $U : \mathcal{D} \times \mathcal{R} \to \mathbb{R}^+$ be a scoring function with sensitivity Δ_U. The* exponential mechanism *sanitizes v to r with probability proportional to* $\exp \frac{\epsilon U(v,r)}{2\Delta_U}$.

3.2.2. Date and Age: Substitution Strategy

Let us first recall that applying a uniform shifting on dates [ULS07, USLS08] is definitively not secure. Indeed this shifting does not modify intervals between any temporal event and these ones are almost unique as shown in [TCC+22].

In [TCC+22], to maintain the chronology of events the following threefold approach is implemented. First, all temporal events of the textual document are normalized into $[e_0, \ldots, e_n]$ an ordered sequence of dates expressed as d-m-y, e_0 being the current date and e_n the oldest one. Next, the sequence $[e_0 - e_1, \ldots, e_{n-1} - e_n]$ of duration between consecutive dates is computed. The input domain is, therefore, $\mathbb{R}^+$ with the absolute value representing the distance. Then, the authors conclude that pertaining to the segmentation of a dataset comprising dates into three distinct categories. These categories were defined based on temporal spans, namely: less than 2 months, less than 2 years, and more than two years. The primary objective of this segmentation was to minimize the sensitivity, which corresponds to the level of introduced noise in the data. However, it is important to note that even within the larger temporal category (more than two years), the amount of introduced noise remained substantial. This is due to the necessity of ensuring the indistinguishability of sanitized dates with a considerable difference in temporal spans, such as 3 years and 80 years.

Therefore, it becomes imperative to conduct a more detailed segmentation of the temporal space. Alternatively, one may consider allowing distinctions between certain dates, especially when the original temporal intervals are significantly apart. The core concept here revolves around the notion that a differential privacy mechanism should not forcibly equate two dates that exhibit substantial temporal disparity. Consequently, the level of privacy protection should be contingent on the distance or dissimilarity between the values of the elements requiring protection. This concept aligns with the framework of ϵ metric-privacy, as introduced in the work by Alvim et al. [ACPP18], which is recalled below.

Definition 4 (ϵ metric-privacy)**.** *A randomized algorithm $\mathcal{A}$ satisfies the ϵ metric-privacy if, for any pair of input values v_1, v_2 in the domain of $\mathcal{M}$ with a metric d and any possible output y of $\mathcal{M}$:*

$$\Pr[\mathcal{M}(v_1) = y] \leq e^{\epsilon.d(v_1,v_2)} \cdot \Pr[\mathcal{M}(v_2) = y]. \tag{3}$$

Instinctively, ϵ metric-privacy safeguards the precision of confidential information: by introducing a metric into the date space, it enables differentiation between an older date (like date of birth) and a recent date (such as the date of an operation last week). Simultaneously, it ensures that two very recent dates (v_1 and v_2) at an extremely small distance will likely produce the same output y.

It has also been demonstrated [CABP13] that the properties of composition and post-processing are possessed by an algorithm that satisfies the ϵ metric-privacy. We are then left to implement a mechanism ensuring this ϵ. metric-privacy property for temporal events. In the same article, it has been be further demonstrated that the mechanism that adds noise following Lap $\left(\frac{1}{\epsilon}\right)$ (i.e. the Laplace distribution centered in 0 and whose scale parameter is $\frac{1}{\epsilon}$) adheres to the ϵ metric-privacy property.

Property 1. *Ensuring ϵ metric-privacy by applying the Laplace mechanism is equivalent to guaranteeing a property of Local Differential Privacy with a Laplace mechanism whose sensitivity is reduced to 1.*

Proof. Let us consider the rate

$$\begin{aligned}\frac{\Pr[\mathcal{M}(v_1) = y]}{\Pr[\mathcal{M}(v_2) = y]} &= \frac{e^{\frac{-|y-v_1|}{b}}}{b} \times \frac{b}{e^{\frac{-|y-v_2|}{b}}} \text{ (where } b = \tfrac{\Delta}{\epsilon}) \\ &= e^{\frac{|y-v_2|-|y-v_1|}{b}} \\ &\leq e^{\frac{|v_2-v_1|\epsilon}{\Delta}}\end{aligned}$$

Achieving ϵ-metric privacy is thus equivalent to verify Local Differential Privacy (LDP) when the sensitivity Δ is 1. □

In all what follows, we consider the rate $\frac{\epsilon}{\Delta}$ to be equal to 1. This approach could potentially be perceived as a privacy vulnerability. In the context of ϵ-LDP, the Δ-amplitude associated with this mechanism would be fixed at 1 day, as opposed to the amplitude corresponding to individual categories, for instance, a substantial range like 100 × 365 days for the largest category. Consequently, specific dates, such as birthdays or medical intervention dates, are prone to substantial modifications. In contrast, age spans covering several decades are less likely to be altered, thereby posing an unsatisfactory outcome. This issue has been addressed in the following manner. First, it is worth noting the following observation: in a medical document containing a statement such as "10 years ago," the intended interpretation is more likely to be "approximately 10 years ago" rather than "on the exact same day 10 years prior." This type of approximation is also applicable when temporal events are expressed in terms of months or weeks. The metric under consideration becomes unit-dependent, employing years (or months, weeks, etc.) for events expressed in the corresponding units. As a result of this adaptation, an individual's age (measured in years) is expected to undergo a modification by a few years, for instance.

It is worth noting that, similar to classical differential privacy approaches, the privacy global ϵ budget is shared among all the elements to be substituted. This sharing can be either uniform or not, and in the absence of specific details, we assume it to be uniform in this context.

Thread Example. In our thread example, there are 2 detected dates (expressed in days), 1 age expressed in years, and 1 location (2 times duplicated), *i.e.* 4 contributions to substitute. Due to the sequential composition theorem [KOV15] of differential privacy, the epsilon budget must be distributed among the various contributions of an individual. Each element will thus consume $\frac{\epsilon}{4}$ of the privacy budget, the last quarter is dedicated to sanitizing location. The date substitution process for this example is detailed in Figure 6. Thus, **40** years, whose unit, and thus whose metric is expressed in years becomes 37 years. Similarly **02/12/2020** and **February 26, 2020** whose unit, and thus whose metric is expressed in days respectively lead to 02/20/2020 and March 01, 2020.

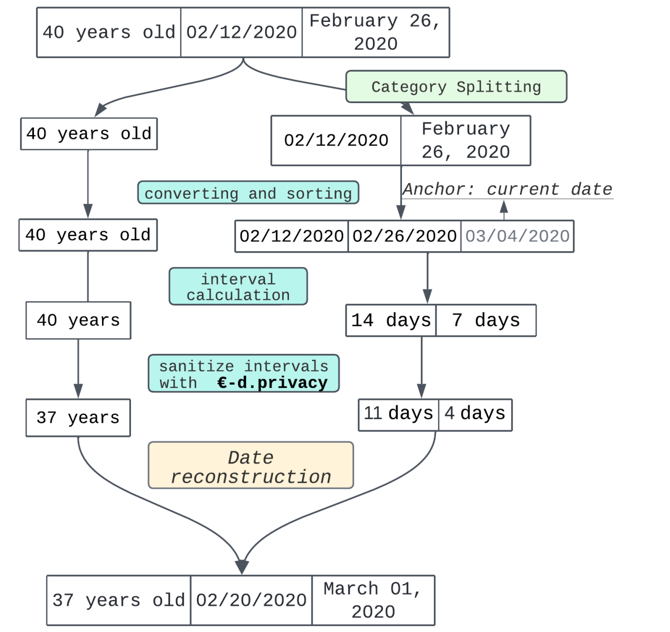


Figure 6: Example of date substitution process on the thread example.

3.2.3. *Geographic Locations: Substitution Strategy*

Geo-indistinguishability [BCP14] is widely accepted as the de facto gold standard [XX15, FS14] for preserving location privacy. This mechanism effectively realizes ϵ metric-privacy (as recalled in definition 4) in the context of locations, which are represented as (x, y) coordinates within $\mathbb{R}^2$.

In the de-identification context, as outlined in [TCC+22], the authors employ geo-indistinguishability to introduce random noise to the coordinates (x, y) of the location to be sanitized, resulting in a new tuple (x', y'). Subsequently, they re-associate the location that is closest to this new tuple. This method successfully safeguards privacy and provides a coherent substitute in the document. However, we contend that it may not fully address the question of the document's utility in a medical context. The concern stems from the fact that two locations that are geographically close to each other may be far apart from a health perspective. For instance, what is the In an example, this could lead to substituting the city of Dijon with the village of Beze, located at a distance of 23 km. However, with the aim of preserving the medical utility of the data, what is the relevance of this substitution when considering that the former has approximately 160,000 inhabitants and rates of cancer and stroke at 182 and 174, respectively, while the latter has about 700 inhabitants with cancer and stroke rates approximately equal to 1?

In the substitution of locations, it is desirable to randomly select among locations that are not only close geographically but also close in a statistical sense, considering factors such as population size, and in a medical sense, considering variables such as the incidence rate of all cancers, the number of strokes, air pollution, radon levels, etc. The key here is to express a distance between locations that incorporates statistical and medical features. Fortunately, many institutional websites provide this local information freely[5][6].

Table 4, within its blue-framed section, provides an excerpt for some cities in France. With features available for each location, it is straightforward to calculate the distance between them, such as the Euclidean distance, and then apply any Local Differential Privacy (LDP) mechanism capable of capturing this distance.

For instance, let's consider a public database $\mathcal{R}$ of N locations, where each location i in $\mathcal{R}$, is represented as a $n+2$ length vector $(x_i, y_i, c_i^1, \dots, c_i^n)$, with (x_i, y_i) being the geographical coordinates, and $(c_i^1, \dots, c_i^n)$ representing features, which are normalized within the range $[0, 1]$. Let d_{ji} be the norm of feature differences between the two locations j and i, *i.e.* $d_{ji} = ||(c_i^1 - c_j^1, \dots, c_i^n - c_i^n)||_2 \in \left[0, \sqrt{n}\right]$.

Let $v_j = [(1, d_{j1}), (2, d_{j2}), \dots, (j, 0), \dots, (N, d_{jN})]$ be the sequence of all distances between j and the other locations. In a practical situation, this sequence can be reduced to the location distances (i, d_{ji}) such that the geographical distance between i and j is lower than a given threshold and the k smallest values of the distances, where values are sorted in ascending order according to d. The sequence $v'_j = [(i_1, d_{ji_1}), \dots, (i_k, d_{ji_k})]$ is this result. The set $\mathcal{R}_j = \{i_1, \dots, i_k\}$ constitutes the possible substitutes for the city j. Notice that $(i_1, d_{ji_1}) = (j, 0)$ since the smallest distance is 0 between j and itself. The retained score function $U : \mathcal{R} \times \mathcal{R} \rightarrow \mathbb{R}^+$ is defined by

$$U(j, i) = \begin{cases} 1 - \frac{d_{ji}}{\sqrt{n}} & \text{if } i\mathcal{R}_j \\ -\infty & \text{otherwise} \end{cases} \tag{4}$$

Notice that $\sqrt{n}$ is a normalization factor allowing the score to belong to $[0, 1]$. This function is public and is not based on any private data. The probability distribution function is thus as follows:

$$P_j = [a.e^{\epsilon U(j, i_1)}, \dots, a.e^{\epsilon U(j, i_k)}, 0, \dots, 0] \tag{5}$$

Where $a = \left(\sum_{i=1}^{k} e^{\epsilon U(j,i)}\right)^{-1}$ is the normalization factor. Notice this mechanism is an adaptation of the centralized exponential mechanism with public data, i.e., without sensitivity. The next section shows it verifies ϵ metric-privacy.

Property 2. *The mechanism defined in previous section verifies ϵ metric-privacy.*

Proof. According to the definition 4, for any y whose probability distribution definition is not null we successively have

$$\frac{\Pr[\mathcal{A}(v_1)=y]}{Pr[\mathcal{A}(v_2)=y]} = \frac{ae^{\epsilon U(v_1,y)}}{ae^{\epsilon U(v_2,y)}} = \frac{e^{\epsilon(1-d(v1,y))}}{e^{\epsilon(1-d(v2,y))}} = e^{\epsilon(d(v_2,y)-d(v_1,y))} \leq e^{\epsilon.d(v_1,v_2)}$$

□

Clearly, the features to be integrated for this step should be defined upstream in collaboration between the medical teams (who have knowledge of the data) and the technical teams.

Thread Example. Using our example with the location "Dijon" and considering features like overall population, cancer incidence rate, and strokes (as shown in blue in Table 4) for a database of all cities in France.

The columns ('distance' and 'scores') represent, respectively, the vector distance (Euclidean distance with normalized features) and the results of the score function U from Dijon to $k = 10$ 'nearby' cities (according to features). After applying the probability distribution function previously detailed, we obtain the normalized distribution illustrated in orange in Table 4. The random draw thus follows this distribution.

According to memoization, all occurrences of the location **Dijon** can be replaced by **Besançon**.

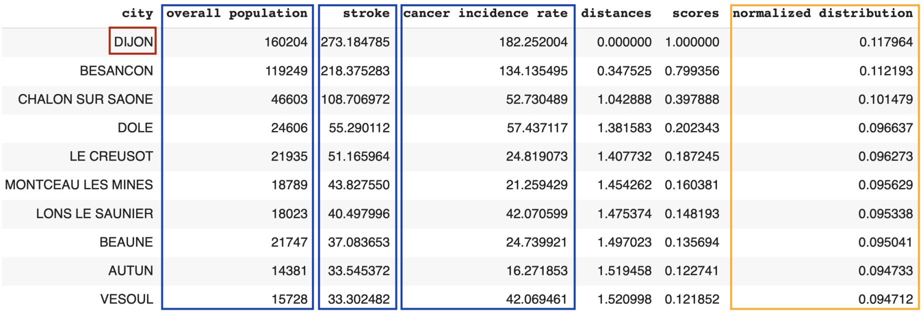
Therefore, the final result of the substitution step would be: "**Mr. Julien, born in Besançon, 37 years old, was admitted to the hospital from 02/20/2020 to March 01, 2020 following a road accident in Besançon.**"

4. Automatic ICD-10 Code Association System

This section presents the ICD-10 automatic code association task and outlines the challenges it poses. Within this context, we present our proposed architecture that address these challenges and the the experiments we have conducted.

[5] https://geodes.santepubliquefrance.fr/

[6] https://www.insee.fr/



city	overall population	stroke	cancer incidence rate	distances	scores	normalized distribution
DIJON	160204	273.184785	182.252004	0.000000	1.000000	0.117964
BESANCON	119249	218.375283	134.135495	0.347525	0.799356	0.112193
CHALON SUR SAONE	46603	108.706972	52.730489	1.042888	0.397888	0.101479
DOLE	24606	55.290112	57.437117	1.381583	0.202343	0.096637
LE CREUSOT	21935	51.165964	24.819073	1.407732	0.187245	0.096273
MONTCEAU LES MINES	18789	43.827550	21.259429	1.454262	0.160381	0.095629
LONS LE SAUNIER	18023	40.497996	42.070599	1.475374	0.148193	0.095338
BEAUNE	21747	37.083653	24.739921	1.497023	0.135694	0.095041
AUTUN	14381	33.545372	16.271853	1.519458	0.122741	0.094733
VESOUL	15728	33.302482	42.069461	1.520998	0.121852	0.094712

Table 4
Example input geographical and epidemiological data used in the sanitization of Dijon city.

4.1. ICD-10 code association task

To ensure accurate long-term follow-up, patient details during their stay in a healthcare facility are typically documented in digital records, constituting the patient's medical record. These records, comprising operative reports, clinical notes, medical correspondence, and other elements, are authored by the attending physicians. In many countries, each patient record is subsequently categorized according to the International Classification of Diseases (ICD), a medical classification system managed by the World Health Organization (WHO) and globally adopted for encoding diseases and other health conditions. This study focuses on the 10th edition of the ICD (ICD-10) [O+92]. Automatically associating ICD-10 codes with medical records has emerged as a significant research challenge in the contemporary medical scientific community [CBS+16, BNKC+18, VNN20, DCC+20, HTC22].

4.2. Challenges

Coding ICD codes involves assigning a set of codes to a given medical record, constituting a multi-label text classification task. However, creating an efficient model for automating this ICD code association is a complex endeavor.

One primary challenge arises from the extensive number of ICD-10 codes, totaling around 140,000 distinct codes, encompassing both procedure and medical codes. Achieving high accuracy by associating any of the 140,000 existing codes with a medical record is unrealistic without access to a massive dataset, significant resources, and considerable time. In practice, ICD-10 code association datasets are notably smaller compared to the comprehensive ICD-10 code set. For instance, the English MIMIC-II and MIMIC-III corpora contain 5,031 and 8,922 different codes (labels), respectively. The substantial number of labels presents a formidable challenge for conventional classification models.

Another significant challenge stems from the size of medical records typically subject to ICD code association. Medical notes often exceed the sequence limit that Transformer architectures can handle, typically set at 512 words. As depicted in Table 5, the average size of medical notes in this study's dataset is approximately 747 words, surpassing the capacity of conventional Transformer (encoders) models.

4.3. Dataset

A patient's stay comprises a series of successive visits to possibly different departments within the hospital. Each department generates a clinical document detailing the patient's experience during their stay in that specific department. These clinical documents are utilized by medical coding specialists for the purpose of associating the relevant ICD-10 codes. Consequently, we obtain a collection of unstructured textual documents that correspond to the comprehensive patient stay, each associated with a set of codes. These clinical documents encompass various types, such as operating reports, discharge letters, external reports, or clinical notes. The resulting dataset, referred to as the ICD-10-HNFC dataset, thus constitutes a repository of groups of medical documents each linked to associated codes. This system is effectively illustrated in Fig. 7.

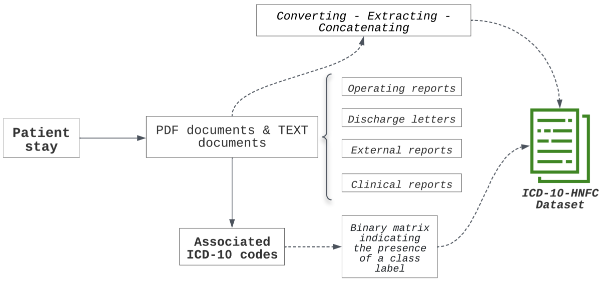

Converting - Extracting - Concatenating
Operating reports
Discharge letters
External reports
Clinical reports
Patient stay
PDF documents & TEXT documents
ICD-10-HNFC Dataset
Associated ICD-10 codes
Binary matrix indicating the presence of a class label

Figure 7: ICD-10-HNFC Dataset Construction.

However, the ICD-10-HNFC dataset remains a private dataset containing sensitive information about individuals. In the interest of privacy, we have applied our whole de-identification approach, proposed in the previous section to this dataset to produce a de-identified dataset that we will refer to as ICD-10-DEID-HNFC. For all these ϵ-metric privacy based algorithms, the privacy parameter ϵ have been set to 1 as well as the value of Delta. The choice of Δ The selection of the Δ value is guided by Property 1, while the choice of the ϵ value is a compromise between utility and privacy.

This process is illustrated in Figure 8. The training process will then be experimented with on the ICD-10-DEID-

	Dataset	Dataset with class reduction
Documents	56000	-
Tokens	41868993	-
Average sequence length	747	-
Total ICD codes	416125	415830
Unique ICD codes	**6160**	**1564**
Codes with less than 10 ex.	3722	523
Codes with 100 ex. or more	641	471

Table 5
Descriptive statistics of ICD-10-HNFC dataset.

HNFC dataset.

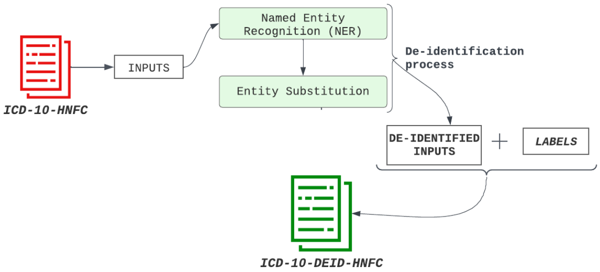


Figure 8: ICD-10-DEID-HNFC Dataset Construction.

Class Reduction

Additionally, we delve into sub-dataset, incorporating the association of the code families. It addresses the challenge of large label set by reducing the codes to the first 3 characters, treated as a family. Instead of considering the raw codes individually, they are grouped into families. This substantial reduction in the number of classes to be handled by the model is evident in Table 5. The "Code with less than 10 examples" category in this Table demonstrates that this reduction not only provides a more manageable number of classes but also increases the frequency of codes in the dataset.

4.4. Model Architecture

This section introduces the various components of the model architecture we have developed and provides justification for the choices made in its design. As previously mentioned, given that we are working with the French language, our approach involves fine-tuning pre-trained transformer-based French models, specifically CamemBERT [MMOS+20] and FlauBERT, to implement the model architecture.

4.4.1. Global Document Representation

As previously highlighted, Transformers face a constraint regarding the maximum number of tokens present in an input sequence. Given that the average size of clinical notes in the ICD-10-HNFC dataset surpasses this limit (747 versus 512, as shown in Table 5), conventional Transformers become impractical. Recently, [DCDE22] provided a summary of available methods for processing long sequences using Transformers, which includes hierarchical Transformers and sparse-attention Transformers. Notably, the *Longformer* model by [BPC20] falls into the category of sparse-attention Transformers and can process up to 4096 tokens per sequence, thereby overcoming this limitation.

However, as of now, there is no pre-trained *Longformer* model available for the French language. Consequently, in this paper, we opt for the hierarchical method to address the challenge posed by the length of clinical notes in the dataset.

Hierarchical Transformers [PZV+19, DCDE22] are an extension of the Transformers architecture. In this approach, a document D is initially divided into segments $[t_0, t_1, \dots, t_{|D|}]$, each containing fewer than 512 tokens (the limit of Transformers). These segments are then encoded independently using a pre-trained Transformers model. Consequently, we obtain a list of segment representations that must be aggregated to form the complete document representation for D. Various aggregation methods can be employed for this purpose. The aggregator may compute the average of the representations of all the segments in the document (mean pooling), take the maximum value of the representations in each dimension across segments (max pooling), or stack the segment representations into a single sequence. The resulting aggregated sequence serves as input to the subsequent layer in the model.

Classification of a Large Number of Labels

To address the challenge of dealing with a large set of labels, given that ICD-10-HNFC contains more than 6,000 codes, we employed the Label-Aware Attention (LAAT) system introduced by [VNN20], similar to the approach in [HTC22]. LAAT involves incorporating the labels into the document representation, allowing it to capture essential text fragments associated with specific labels.

Let H denote the stacked representation of an input sequence. Initially, a label-wise attention weight matrix Z is computed as follows:

$$Z = tanh(VH)$$

$$A = softmax(WZ)$$

where V and W are linear transforms. The i^{th} row of A represents the weights of the i^{th} label. The softmax function is performed for each label to form a distribution over all tokens. Then, the matrix A is used to perform a weighted-sum of H to compute the label-specific document representation:

$$D = HA^T$$

The i^{th} row of D represents the document representations for the i^{th} label. Finally, D is used to make predictions by computing the inner product between each row of D and the related label vector.

In this paper, several architectures were experimented with, including the model without long sequence processing, the model with long sequence processing using max/mean pooling, and the model with Label-Aware Attention (LAAT). The overall architecture is illustrated in Fig. 9.

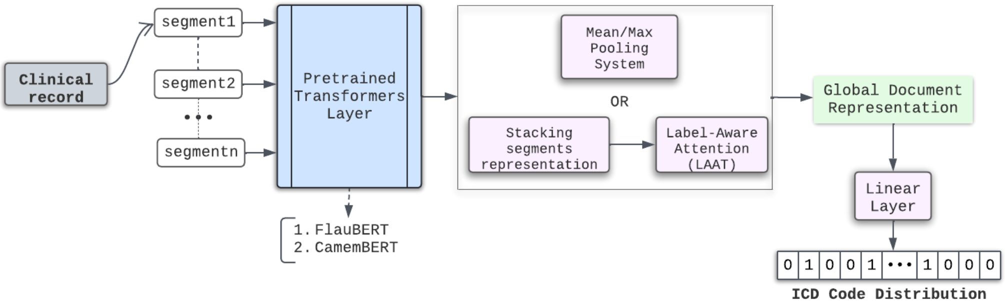


Figure 9: Architecture of ICD-10 code association system.

5. Effect of de-identification on medical utility

In this section, we will evaluate the utility preservation of the de-identification method presented in Section 3.

To measure the usefulness of de-identification, we use the machine learning system we just addressed in section 4: ICD-10 code association. This task is relevant for measuring the usefulness of de-identification, because the data used not only contain the sensitive information we are seeking to de-identify, such as names, geographical locations, ages, dates, etc., but also some of this information may influence the medical analysis of the document (e.g. the patient's age may direct the analysis towards a certain code category, or patients from the same location marked by a natural factor may present the same pathologies, etc.). The usefulness of de-identification will therefore be measured by the association performance or errors with a learning technique before and after the de-identification process. The construction of the datasets is detailed below.

5.1. Methodology

This section presents the methodological approach used to evaluate the utility of de-identification strategies. We begin by introducing the de-identification strategies used to construct the various datasets. Subsequently, we present the machine learning system that facilitated the evaluation of these different datasets.

5.1.1. De-identified datasets ICD-10-DEID-HNFC and ICD-10-TAG-HNFC

We consider the original data, represented by the dataset ICD-10-HNFC, which is described in Section 4.3. This dataset is the input to generate the de-identified dataset ICD-10-DEID-HNFC following differential privacy based methods presented in Section 3.2 aiming at establishing our automatic association system for ICD-10 codes. Notice that as previously justified, the rate $\frac{\epsilon}{\Delta}$ is 1. The process of creating this de-identified datasets is illustrated in Figure 8.

A second dataset ICD-10-TAG-HNFC has been generated by replacing each detected entity in ICD-10-HNFC by its associated entity label: the city name "DIJON" in the document is thus be replaced by the entity "LOC" (for location). This de-identification approach is the most robust against potential re-identification attacks. However, by removing the values of the entities, they cannot be exploited for classification purposes.

5.1.2. ICD-10 code Association system

After creating the various datasets, we implemented the ICD-10 code association system inspired by [HTC22] based on the architecture detailed in Section 4.4 to address the challenges mentioned earlier. The supervised learning process (including training, validation, and evaluation), is then applied to the three datasets: ICD-10-HNFC, ICD-10-DEID-HNFC, and ICD-10-TAG-HNFC. This process is illustrated in Figure 10.

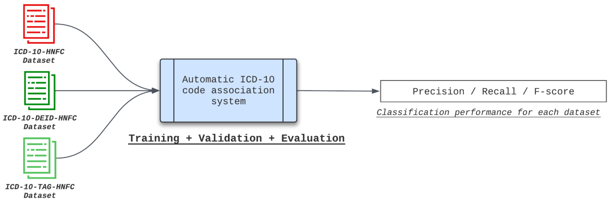

ICD-10-HNFC Dataset
ICD-10-DEID-HNFC Dataset
ICD-10-TAG-HNFC Dataset
Automatic ICD-10 code association system
Training + Validation + Evaluation
Precision / Recall / F-score
Classification performance for each dataset

Figure 10: Approach to evaluate the utility of de-identification.

6. Experiments

This section presents the evaluations of the paper. Initially, we measure the utility of de-identification with a CIM code association system inspired by [HTC22] on datasets with various levels of de-identification. Then, we present the results of different architectures developed within the scope of this work on the original dataset.

The evaluation of our models is performed using commonly used classification performance measures: Precision, Recall, and F_1-score. The micro-average system is employed for aggregating the performances.

Dataset	Labels	**Precision**	**Recall**	**F_1-score**
ICD-10-HNFC	**6160**	**0.47**	**0.46**	**0.47**
ICD-10-DEID-HNFC		0.44	0.43	0.44
ICD-10-TAG-HNFC		0.43	0.41	0.42

Table 6
Evaluation of different datasets on the same validation dataset.

6.1. Experiments on the usefulness of de-identification

Regarding the evaluation, we have chosen the main task, which involves classifying all the codes present in the dataset (6160 labels, as indicated in the dataset's descriptive table in 5). We apply the same performance criteria as those described above, namely precision, recall, and the F_1 score. The experiments conducted on the three datasets with the same instance of input data using the same parameters. The results obtained are presented in Table 6.

We observe that the results obtained from the evaluations show disparities among the datasets, despite using the same learning process and validation data. These variations confirm, even without considering the results, that de-identification has an impact on the medical document analysis, specifically regarding the association of ICD-10 codes.

The highest performance among the three datasets is obtained from the evaluation of the original corpus ICD-10-HNFC. This illustrates how the de-identification process leads to a deterioration of the document's utility. The evaluation on the ICD-10-TAG-HNFC corpus resulted in the lowest F_1 score. Considering the evaluations conducted on the ICD-10-DEID-HNFC and ICD-10-TAG-HNFC corpora, which correspond to the de-identification proposed in this paper and partial de-identification, respectively, it is clear that partial de-identification is the one that most negatively affects the document's utility. The percentage gap between the original corpus ICD-10-HNFC and the de-identified corpus ICD-10-TAG-HNFC is approximately 12%. The de-identification process described in Section 3 reduces this gap to 6.8%. This reinforces the idea that the question of medical utility in the context of de-identification is perfectly justified.

6.2. Experiments on original dataset: ICD-10-HNFC

In this section, we present the experiments conducted on the original dataset using various architectures. We compare the outcomes with recent works, such as PLM-ICD [HTC22] and CNN [DCC+20].

6.2.1. Models

Our initial experiments were conducted on the ICD-10-HNFC dataset with class reduction (1564 labels), as outlined in Table 7. These experiments encompassed all the architectures developed in this paper, as listed in Table 7. Subsequently, we trained another model on the complete ICD-10-HNFC dataset (6160 labels) using the architectures that achieved the highest F_1-score in the previous experiment. The results are presented in Table 7. These findings affirm the impact of the different components within our architectures. In summary, the LAAT approach outperforms the hierarchical methods, which, in turn, surpass the base truncated model.

Models	Labels	Precision	Recall	F_1-score
FlauBERT (512 tokens)	1564	0.48	0.31	0.38
Hierarchical Mean FlauBERT		0.54	0.39	0.45
Hierarchical Max FlauBERT		0.53	0.40	0.46
FlauBERT + LAAT		**0.57**	0.51	0.54
CamemBERT + LAAT		0.56	**0.53**	**0.55**
FlauBERT + LAAT	6160	0.41	0.43	0.42
CamemBERT + LAAT		0.52	0.4	**0.45**

Table 7
ICD-10 association results of the different architectures on the validation ICD-10-HNFC dataset.

Models	Language	Dataset	Labels	F_1-score
PLM-ICD [HTC22]	*English*	*MIMIC 2*	*5,031*	*0.5*
		MIMIC 3	*8,922*	***0.59***
[DCC+20]	*French*	*[DCC+20]*	*6,116*	*0.39*
			1,549	*0.52*
PROPOSAL	French	ICD-10-HNFC	6,160	**0.45**
			1,564	**0.55**
[DCC+20]			6,160	0.27
			1,564	0.35

Table 8
Results comparison with the previous work on ICD-10 Association. The state of the art works with their results are in *italic*. The experiments done in this paper with ICD-10-HNFC dataset are presented in the other part. The highest scores in each part in relation to the number of labels are marked in **bold**.

6.2.2. Analysis

Table 8 showcases the model with the highest F_1-score from this paper alongside the results of previous work on ICD-10 code association. Comparing these results is challenging due to the variation in evaluation datasets, and English works may benefit from specialized models such as ClinicalBERT [AMB+19]. As a French baseline, we implemented and trained the model proposed in [DCC+20] on the ICD-10-HNFC dataset. The results are juxtaposed with our proposal.

Our model significantly outperforms the classification method used in [DCC+20]. On the same validation dataset, with class reduction (1564 labels), the F_1-score improves from 0.35 (model proposed in [DCC+20]) to 0.55 with our proposal, representing a 57% improvement. With the raw codes (6160 labels), the F_1-score improves from 0.27 to 0.45, indicating a 66.6% improvement. The variance in scores compared to PLM-ICD results may be attributed to the use of a context-specific (medical) Transformers with a vocabulary more tailored to the content of medical documents.

7. Discussion

An observation potentially raised regarding this research could pertain to the utilization of models deemed "outdated" in comparison to those relying on Large Language Models . Nevertheless, this can be mitigate by considering two fundamental points. Firstly, our study focused on the specific domain of medical French, for which, to the best of our knowledge, there exists no appropriate substantiation of LLM. Secondly, due to confidentiality imperatives, all stages of development and experimentation were conducted within the hospital site. This particularity resulted in a complexification of the process, given that the potentially accessible hardware significantly differs from that typically available in a conventional computing environment, and data access is subject to stringent regulations aimed at facilitating our approach. These constraints unquestionably extended the required duration for project completion, a consideration particularly detrimental in the current temporal context.

8. Conclusion

This paper addresses the de-identification of medical data for medical analyses within the framework of scientific research. Firstly, we tackle the de-identification task itself, which encompasses the Named Entity Recognition (NER) task and the generation of substitutes. For the Named Entity Recognition, an existing, comprehensive but imperfect de-identification approach was used to internally construct a new and substantial dataset of medical records (approximately 15,000 sentences). Leveraging this large labeled dataset, deep learning was implemented with a Transformer-based architecture. The results obtained represent the state-of-the-art in the French language and are as precise as those achieved in English.

To substitute the detected sensitive information, we integrate the notion of medical utility through the differential privacy mechanism based on a metric (ϵ metric-privacy), mainly at the level of attributes that can impact medical analysis, such as temporal data (dates and ages) and geographical locations. To evaluate the proposed de-identification strategies, we place this work in a direct application context, which is the association of ICD-10 codes.

We start by addressing the scientific challenges of ICD-10 code association, which is a multi-label text classification task, namely the significant number of ICD-10 codes and the size of medical data. In this paper, we propose the most accurate model to date for the automatic association of ICD-10 codes in the French language. Subsequently, we used this machine learning system to evaluate our de-identification approach. This evaluation confirmed the preservation of utility in our approach compared to conventional de-identification methods (e.g., substituting attributes with their labels).

Even if the presented approach meets the requirements regarding privacy protection by incorporating differential privacy, resulting datasets are not immune from a re-identification or membership attack on the dataset. Therefore, even de-identified datasets cannot be disseminated. The question of the risk of re-identification of de-identified documents remains complex. An in-depth analysis of vulnerabilities and potential attacks against the algorithms proposed in this paper is necessary an is planed as a future work to provide a more precise answer to this question.

Finally, hospitals publishing ICD-10 codes generate temporal statistics for institutions and receive funding based on care provided, with codes serving as a summarized version. The objective can thus be seen as more medically administrative than purely medical. On the opposite, using personal data, even de-identifid one in medical research falls under the application of the French law Jardé[7], which is much stricter and takes longer to implement. The theoretical and practical approaches taken in this article, both in de-identification and classification, can also serve in such contexts.

This study empirically demonstrated that the utility of the de-identification process decreased by 12% (in terms of a classification score) when replacing each entity with its TAG. It decreased by 6.8% when using an ϵ-LDP approach, where the ratio $b = \frac{\epsilon}{\Delta}$ is 1. Formally, substituting an entity with its TAG corresponds to de-identification with $b = 0$, and not modifying the attributes corresponds to $b = \infty$. Between these extreme values, we plan to evaluate how the classification score evolves as a function of the ratio b. As application future work, we plan to detect treatment pathways disrupted by healthcare-associated infections

References

[ABB+19] A. Akbik, Tanja Bergmann, Duncan A. J. Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL*, 2019.

[ABBZ17] Ankur Agarwal, Christopher Baechle, Ravi Behara, and Xingquan Zhu. A natural language processing framework for assessing hospital readmissions for patients with copd. *IEEE journal of biomedical and health informatics*, 22(2):588–596, 2017.

[ACABX22] Héber H Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates. *Digital Communications and Networks*, 2022.

[ACPP18] Mário S Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. Metric-based local differential privacy for statistical applications. *arXiv preprint arXiv:1805.01456*, 2018.

[AHL+22] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*, 2022.

[AMB+19] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[AND+19] Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert. In *CLEF (Working Notes)*, pages 1–15, 2019.

[BAC+21] Loick Bourdois, Marta Avalos, Gabrielle Chenais, Frantz Thiessard, Philippe Revel, Cédric Gil-Jardiné, and Emmanuel Lagarde. *De-identification of Emergency Medical*

[7] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000025441587

Records in French: Survey and Comparison of State-of-the-Art Automated Systems, volume 34. LibraryPress@UF, May 2021.

[BBBK11] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

[BCP14] Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Optimal geo-indistinguishable mechanisms for location privacy. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 251–262, 2014.

[BGJM16] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016.

[BNKC$^+$18] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018.

[BPC20] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[BYC13] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

[CABP13] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer, 2013.

[CBS$^+$16] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[CCL$^+$20] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. Hypercore: Hyperbolic and co-graph representation for automatic icd coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, 2020.

[CKG$^+$19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[CM18] I Glenn Cohen and Michelle M Mello. Hipaa and protecting health information in the 21st century. *Jama*, 320(3):231–232, 2018.

[CVFW23] Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. Tem: High utility metric differential privacy on text. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 883–890. SIAM, 2023.

[DBKMR05] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005.

[DCC$^+$20] Clément Dalloux, Vincent Claveau, Marc Cuggia, Guillaume Bouzillé, and Natalia Grabar. Supervised learning for the icd-10 coding of french clinical narratives. In *MIE 2020-Medical Informatics Europe conference-Digital Personalized Health and Medicine*, pages 1–5, 2020.

[DCDE22] Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. Revisiting transformer-based models for long document classification. *arXiv preprint arXiv:2204.06683*, 2022.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DJW13] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[DKY17] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

[DLN$^+$14] Louise Deleger, Todd Lingren, Yizhao Ni, Megan Kaiser, Laura Stoutenborough, Keith Marsolo, Michal Kouril, Katalin Molnar, and Imre Solti. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *Journal of biomedical informatics*, 50:173–183, 2014.

[DLUS16] Franck Dernoncourt, Ji Lee, Ozlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association : JAMIA*, 24, 06 2016.

[EFM$^+$19] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[EU16] EU. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016.

[FBDD20] Oluwaseyi Feyisetan, Borja Balle, Tom Diethe, and Thomas Drake. Calibrating mechanisms for privacy preserving text analysis. In *PrivateNLP@ WSDM*, pages 8–11, 2020.

[FGG$^+$07] M Fabian, Kasneci Gjergji, WEIKUM Gerhard, et al. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International world wide web conference, WWW*, pages 697–706, 2007.

[FS14] Kassem Fawaz and Kang G Shin. Location privacy protection for smartphone users. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 239–250, 2014.

[GGN15] Cyril Grouin, Nicolas Griffon, and Aurélie Névéol. Is it possible to recover personal health information from an automatically de-identified corpus of french ehrs? In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 31–39, 2015.

[Han21] Ridewaan Hanslo. Deep learning transformer architecture for named entity recognition on low resourced languages: State of the art results. *CoRR*, abs/2111.00830, 2021.

[HHD$^+$20] Tzvika Hartman, Michael Howell, Jeff Dean, Shlomo Hoory, Ronit Slyper, Itay Laish, Oren Gilon, Danny Vainstein, Greg Corrado, Katherine Chou, Ming Po, Jutta Williams, Scott Ellis, Gavin Bee, Avinatan Hassidim, Rony Amira, Genady Beryozkin, Idan Szpektor, and Yossi Matias. Customization scenarios for de-identification of clinical notes. *BMC Medical Informatics and Decision Making*, 20, 01 2020.

[HR18] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[HSDF21] Fadi Hassan, David Sánchez, and Josep Domingo-Ferrer. Utility-preserving privacy protection of textual documents via word embeddings. *IEEE transactions on knowledge and* *data engineering*, 35(1):1058–1071, 2021.

[HTC22] Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. Plm-icd: automatic icd coding with pretrained language models. *arXiv preprint arXiv:2207.05289*, 2022.

[JFHS22] Sara Jordan, Clara Fontaine, and Rachele Hendricks-Sturrup. Selecting privacy-enhancing technologies for managing health data use. *Frontiers in Public Health*, 10:814163, 2022.

[JPS$^+$16a] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[JPS$^+$16b] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[KCG$^+$23] Marianne Abi Kanaan, Jean-François Couchot, Christophe Guyeux, David Laiymani, Talar Atéchian, and Rony Darazi. A methodology for emergency calls severity prediction: From pre-processing to bert-based classifiers. In Ilias Maglogiannis, Lazaros S. Iliadis, John MacIntyre, and Manuel Domínguez, editors, *Artificial Intelligence Applications and Innovations - 19th IFIP WG 12.5 International Conference, AIAI 2023, León, Spain, June 14-17, 2023, Proceedings, Part I*, volume 675 of *IFIP Advances in Information and Communication Technology*, pages 329–342. Springer, 2023.

[KOV15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[LBD$^+$23] Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. Drbert: A robust pre-trained model in french for biomedical and clinical domains, 2023.

[LCK$^+$21] Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953, 2021.

[LH17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[LOG$^+$19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[LPS$^+$21] Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, 2021.

[LTWC17] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75, 06 2017.

[LVF$^+$19] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*, 2019.

[LYK$^+$20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[LYZ$^+$23] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023.

[MDS23] Soha Sadat Mahdi, Nikos Deligiannis, and Hichem Sahli. A review of deep learning methods for automated clinical coding. In *2023 15th International Conference on Computer and Automation Engineering (ICCAE)*, pages 35–39. IEEE, 2023.

[MHL$^+$17] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*, pages 2285–2294. PMLR, 2017.

[MMOS$^+$20] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[MWD$^+$18] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*, 2018.

[NKK$^+$18] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. doccano: Text annotation tool for human, 2018. Software available from https://github.com/doccano/doccano.

[NRG$^+$18] Aurélie Névéol, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, Laszlo Pelikan, Lionel Ramadier, Grégoire Rey, and Pierre Zweigenbaum. Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In *CLEF (Working Notes)*, pages 1–18, 2018.

[NRR$^+$13] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175, 2013. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

[O$^+$92] World Health Organization et al. Icd-10. international statistical classification of diseases and related health problems: Tenth revision. 1, 1992.

[PAL$^+$22] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[PdGS21] Marco Polignano, Marco de Gemmis, and Giovanni Semeraro. Comparing transformer-based NER approaches for analysing textual medical diagnoses. In Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi, editors, *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 818–833. CEUR-WS.org, 2021.

[Pol] Jean-Baptiste Polle. camembert-ner: model fine-tuned from camembert for ner task. `https://huggingface.co/Jean-Baptiste/camembert-ner`.

[PTM$^+$23] Wanchana Ponthongmak, Ratchainant Thammasudjarit, Gareth J McKay, John Attia, Nawanan Theera-Ampornpunt, and Ammarin Thakkinstian. Development and external validation of automated icd-10 coding from discharge summaries using deep learning approaches. *Informatics in Medicine Unlocked*, 38:101227, 2023.

[PZV+19] Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE, 2019.

[RWC+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[SBS+20] Alessandro Scardoni, Federica Balzarini, Carlo Signorelli, Federico Cabitza, and Anna Odone. Artificial intelligence-based tools to control healthcare associated infections: a systematic review of the literature. *Journal of infection and public health*, 13(8):1061–1077, 2020.

[SGCP20] Stefano Silvestri, Francesco Gargiulo, Mario Ciampi, and Giuseppe De Pietro. Exploit multilingual language model at scale for icd-10 clinical text classification. *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7, 2020.

[SKW+22] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*, 2022.

[SRU13] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, 2013.

[SUK+15] Amber Stubbs, Özlem Uzuner, Christopher Kotfila, Ira Goldstein, and Peter Szolovits. Challenges in synthesizing surrogate phi in narrative emrs. In *Medical data privacy handbook*, pages 717–735. Springer, 2015.

[SVR+11] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[Swe02] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[SXH+17] Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*, 2017.

[TCC19] Shang-Chi Tsai, Ting-Yun Chang, and Yun-Nung Chen. Leveraging hierarchical category knowledge for data-imbalanced multi-label diagnostic text understanding. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 39–43, 2019.

[TCC+22] Yakini Tchouka, Jean-François Couchot, Maxime Coulmeau, David Laiymani, Philippe Selles, and Azzedine Rahmani. De-identification of french unstructured clinical notes for machine learning tasks. *CoRR*, abs/2209.09631, 2022.

[TCL23a] Yakini Tchouka, Jean-François Couchot, and David Laiymani. An easy-to-use and robust approach for the differentially private de-identification of clinical textual documents. In Federico Cabitza, Ana L. N. Fred, and Hugo Gamboa, editors, *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2023, Volume 5: HEALTHINF, Lisbon, Portugal, February 16-18, 2023*, pages 94–104. SCITEPRESS, 2023.

[TCL+23b] Yakini Tchouka, Jean-François Couchot, David Laiymani, Philippe Selles, and Azzedine Rahmani. Automatic ICD-10 code association: A challenging task on french clinical texts. In João Rafael Almeida, Myra Spiliopoulou, José Alberto Benítez-Andrades, Giuseppe Placidi, Alejandro Rodríguez González, Rosa Sicilia, and Bridget Kane, editors, *36th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2023, L'Aquila, Italy, June 22-24, 2023*, pages 91–96. IEEE, 2023.

[TKB+18] Nastassia Tvardik, Ivan Kergourlay, André Bittar, Frédérique Segond, Stefan Darmoni, and Marie-Hélène Metzger. Accuracy of using natural language processing methods for identifying healthcare-associated infections. *International Journal of Medical Informatics*, 117:96–102, 2018.

[TST22] Nischay Bikram Thapa, Sattar Seifollahi, and Sona Taheri. Hospital readmission prediction using clinical admission notes. In David Abramson and Minh Ngoc Dinh, editors, *ACSW 2022: Australasian Computer Science Week 2022, Brisbane, Australia, February 14 - 18, 2022*, pages 193–199. ACM, 2022.

[ULS07] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563, 2007.

[USLS08] Özlem Uzuner, Tawanda C. Sibanda, Yuan Luo, and Peter Szolovits. A de-identifier for medical discharge summaries. *Artificial intelligence in medicine*, 42 1:13–35, 2008.

[VNN20] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for icd coding from clinical text. *arXiv preprint arXiv:2007.06351*, 2020.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WXY+19] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 638–649. IEEE, 2019.

[XX15] Yonghui Xiao and Li Xiong. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1298–1309, 2015.

[XX18] Pengtao Xie and Eric Xing. A neural architecture for automated icd coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, 2018.

[YLZ+20] Mengmeng Yang, Lingjuan Lyu, Jun Zhao, Tianqing Zhu, and Kwok-Yan Lam. Local differential privacy and its applications: A comprehensive survey. *arXiv preprint arXiv:2008.03686*, 2020.

[ZCC+21] Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. Automatic icd coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5948–5957, 2021.